A ranking tool for ”category killer” microbial biobanks

Martin Boutroux, Adriana Chiarelli, Mariana L Ferrari, Olivier Chesneau, D. Clermont, Fay Betsou

A ranking tool for “category killer” microbial biobanks

Boutroux, Martin[1#*]; Chiarelli, Adriana[1#]; Ferrari, Mariana L. [1]; Chesneau, Olivier[2]; Clermont[2], Dominique; Betsou, Fay[1§]

[1]Institut Pasteur, Université Paris Cité, Biological Resource Center of Institut Pasteur – Project Management Office, F-75015, Paris, France

[2]Institut Pasteur, Université Paris Cité, Biological Resource Center of Institut Pasteur – Collection de l’Institut Pasteur, F-75015, Paris, France

[#]Equal contribution

*Current address: River Ecosystems Laboratory, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

[§]Corresponding author: fay.betsou@pasteur.fr

Keywords

Ranking; culling; microbial biobanks; WGS; bacterial strains

ORCID IDs

Martin Boutroux: 0000-0003-0180-7140

Adriana Chiarelli: 0000-0002-3561-0576

Mariana L. Ferrari: 0000-0002-0441-9242

Olivier Chesneau: 0000-0002-9718-0444

Dominique Clermont: 0000-0002-6018-2462

Fay Betsou: 0000-0002-0558-4653

Abstract

Microbial biobanks preserve and provide microbial bioresources for research, training, and quality control purposes. They ensure the conservation of biodiversity, contribute to taxonomical research, and support scientific advancements. Microbial biobanks can cover a wide range of phylogenetic and metabolic diversity ("category killers"), or focus on specific taxonomic, thematic, or disease areas. The strategic decisions about strain selection for certain applications or for the biobank culling necessitate a method to support prioritization and selection.

Here, we propose an unbiased scoring approach based on objective parameters to assess, categorize, and assign priorities among samples in stock in a microbial biobank. We describe the concept of this ranking tool and its application to identify high priority strains for whole genome sequencing with two main goals: (i) genomic characterization of quality control, reference, and type strains; (ii) genome mining for the discovery of natural products, bioactive and antimicrobial molecules, with focus on human diseases. The general concept of the tool can be useful to any biobank and for any ranking or culling needs.

Introduction

Microbes are found abundantly in various ecosystems, and can be a source of plant, animal, and human diseases. However, they are also a source of useful applications in diverse sectors, including the production of bioactive compound[1], food industry[2], and bioremediation[3]. By collecting, authenticating, maintaining, and distributing microbial resources, microbial biobanks ensure the conservation of biodiversity, contribute to taxonomical research, and support scientific advancements. Strategically, microbial biobanks can either expand in breadth, covering a wide range of phylogenetic and metabolic diversity ("category killers"), and/or focus on specific taxonomic, thematic, or disease areas[4]. Given that less than 1% of microbial biodiversity has been cultured to date[5], expanding in breadth allows microbial biobanks to support the preservation of microbial biodiversity, taxonomic research, and biotechnological and medical applications. Alternatively, focusing on specific thematic or disease areas involves preserving specific groups of microbes and providing expertise and support in those specialized fields. Indeed, biobanking, encompassing accurate characterization of microbial specimens, is essential for epidemiological studies, identification of etiological agents of infectious diseases, understanding pathogenicity and virulence, development of diagnostic assays, surveillance of resistance and mutations, and prediction of clinical outcomes[6,7].

However, many microbial biobanks face ongoing challenges of limited funding, staff shortages, inadequate facilities, and weak communication strategies[8,9]. This puts pressure on microbial biobanks

and emphasizes the need for selective bioresource preservation to avoid duplication and dispersion. In this sense, the concept of "key strain" has been proposed to prioritize strain acquisition. This concept is based on criteria like phylogenetic uniqueness, availability of sequenced genomes, inclusion of reference and test strains, samples from unexplored environments, or association with relevant diseases[10].

Similar to clinical biobanks, which commonly experience a rate of sample utilization of approximately 10% and face challenges related to culling[11,12], culling unusable or barely usable strain specimens is often neglected but is crucial to optimize microbial biobank performance[13]. Factors such as insufficient consideration of actual scientific needs, insufficient evaluation and documentation of quality of strains or associated data, or access restrictions imposed by the Nagoya Protocol, can potentially impact the utilization rate of samples in microbial biobanks.

The implementation of a well-defined strategy for stock ranking and culling can support biobank sustainability[14,15]. This strategic approach involves maintaining an up-to-date inventory, which not only aids in identifying higher-value samples but also facilitates the elimination of low-quality specimens. Furthermore, the strategic transfer of samples to more suitable infrastructures becomes integral to optimizing resource allocation.

Crucially, the development and application of such a strategy go beyond mere sample culling. By effectively ranking samples based on objective parameters such as quality control, utilization history, and alignment with institutional priorities, the proposed unbiased scoring approach offers a systematic means to identify high-priority strains for in-depth characterizations, such as whole genome sequencing (WGS). This ensures that resources are directed toward samples with the greatest potential scientific impact, thereby enhancing the overall cost-effectiveness and efficiency of microbial biobanks.

We describe the concept of the tool and its application to identify high priority strains for WGS with two main goals: (i) genomic characterization of quality control, reference, and type strains; (ii) genome mining for the discovery of natural products, bioactive and antimicrobial molecules, with focus on human diseases. The general concept of the tool can be useful to any biobank and for any ranking needs.

Methodology

The CIP (Collection of Institut Pasteur) houses about 25,000 prokaryotic strains, mainly bacteria, encompassing more than 1,000 genera and 5,000 species (Figure 1) from various sources of isolation, and is part of the CRBIP (Biological Resource Center of Institut Pasteur, https://www.pasteur.fr/en/public-health/biological-resource-center). Over 13,000 of these strains have undergone authentication, genotypic, and/or phenotypic quality control, and are publicly accessible

through the CIP catalog (https://catalogue-crbip.pasteur.fr/), while the rest consists of microorganisms deposited mainly by former research groups from Institut Pasteur, and whose timely authentication has not been possible. Only 2,300 strains have been thoroughly characterized by WGS until now and a new CRBIP strategic plan (ref. CRBIP Strategic plan 2023-2032, unpublished, internal document), which is aligned with the Institut Pasteur strategic plan, foresees the possibility to sequence and characterize taxonomically up to 3,000 additional strains to maximize the chances of identifying novel genetic elements and biotechnologically relevant traits through genomic analyses.

The ranking tool is based on an unbiased scoring approach and was developed to enable CIP to achieve this selection, but the general concept can be useful to any biobank and for any ranking needs. A prioritization tool requires definition of critical attributes and assignment of objective or conventional values to the different attributes. In the case of microbial collections, different metadata and analytical data attributes can be used to perform this evaluation. The strain metadata include information on strain's origin, country and date of isolation, legal status relative to the Nagoya Protocol, presence in other biobanks, or distribution rates to users. The analytical data correspond to quality control or characterization assays that have been conducted on the strains, such as confirmation of the identity through mass spectrometry and/or targeted gene analysis or WGS. Those data are usually stored in a BIMS (Biobank Information Management System) that is accessed programmatically. The CIP BIMS data was used in two ways to establish the ranking and culling tool:

(1) Definition of key ranking parameters.

(2) Definition of groups of strategic interest.

1. Definition of key ranking parameters

The parameters aim to evaluate the availability and inherent value of metadata and analytical data attributes within biobanks. A list of relevant parameters was compiled, and a score was attributed to the different qualitative or quantitative values each of those parameters could take (Table 1). A cumulative numeric score could then be calculated to assess the overall usefulness of a strain or a group, and compare it to other strains or groups.

In the case of the CIP, 12 parameters corresponded to strain metadata (Table 1, "Metadata score"). The "Historical period" parameter distinguished strains from different periods: those before 1950 represent the microbial world prior to the massive use of antibiotics; strains before 2014 are not in the scope of Nagoya Protocol (NP) according to the EU Regulation 511[16]; and strains isolated after 2014 are in the scope of NP. The latter are also considered of higher quality in terms of viability and molecular integrity, since *de facto* they have been acquired and stored after 2014. The parameter "Exclusivity" identified strains absent from other microbial biobanks, making them unique to the CIP. Other parameters were related to the availability of metadata that are essential for the usability of a strain as a reference strain or in epidemiological studies, such as country of origin, source, and year of isolation ("Associated data"), and quality control strains included in the World Data Centre for

Microorganisms (WDCM) standard lists ("WDCM standard"). Finally, the complexity of handling ("Collection replenishment") or storing ("Storage temperature") a strain in the laboratory and its popularity among users ("Distribution rate") were assessed, among others.

For the analytical evaluation, four parameters were used (Table 1, "Analytical score"). They corresponded to the types of assays that may have been performed on each strain, including API (Analytical Profile Index) test strips, antibiograms, MALDI-TOF MS, and 16S rRNA or housekeeping gene sequencing. These values indicate the extent of already performed characterization of a strain.

Additional parameters were considered but were not used for the final scoring because of a lack of information readily available in our BIMS. Those parameters dealt with metadata related to legal issues [PIC (Prior Informed Consent from country of origin), MAT (Mutually Agreed Terms), legal requirements for maintenance], and analytical data like phenotypic testing by Biolog plates. Since the objective of the application of the tool was to select strains for WGS, previously performed WGS was not used as a ranking parameter. That or any other parameters could be used if the ranking had a different objective.

2. Definition of groups of strategic interest

Given the high number of CIP strains, it would have been impractical to assign a priority score to each individual strain. Instead, our strategy was to categorize the strains into groups of interest. Given that the CIP is not limited to strains related to clinical research but rather hosts a wide diversity of bacteria and even archaea ("category killer"), this grouping strategy allowed us to delineate several subcollections with specific features. The ranking scores could then be calculated for those subcollections.

Eight groups, designed to represent high taxonomic diversity while maintaining a small number of groups, were created to accommodate the CIP strains (Figure 2):

- One group contained the small number of archaeal strains present at CIP, taxonomic outliers of the overall bacterial collection.
- Three groups relied on publicly available lists of bacteria of special strategic interest: (i) a global priority pathogens list published by the World Health Organization (WHO)[17]; (ii) a list of foodborne zoonotic bacterial species of high priority involved in the emergence and spread of antimicrobial resistance (AMR) through the food chain as defined by European Food Safety Authority (EFSA)[18]; and (iii) a list of quality control strains from the list established by the World Data Centre of Microorganisms (WDCM) (https://refs.wdcm.org/refs/).
- Two other groups relied on specific metadata values: (i) a group of type strains, subdivided into those previously sequenced by other biobanks or laboratories, and those never-sequenced; and (ii) a group based on the source of isolation of the strains, distinguishing between "soil"-isolated bacteria from "eukaryotes"-isolated bacteria (strains isolated from humans or other animals).

- Two groups were based on strain specific characteristics: (i) extremophilic strains, with a distinction made between literature or analytical evidence-based extremophilic strains and potential extremophilic strains on the basis of their metadata; (ii) strains producing special elements, with subgroups of strains producing "pigment" (or being "luminescent"), "indole" (according to[19]), or natural compound producers listed in the NPAtlas database (https://www.npatlas.org/)[20].
- Considering the groups and subgroups based on the eight features described above, a total of 13 different (sub)groups were defined. Groups and subgroups were treated the same, therefore they are all qualified as groups thereafter. A 14th group contained all strains belonging to none of the other groups.

The lists of strains belonging to each group were extracted from our PostgreSQL database with SQL queries and Python scripts (available on GitHub at github.com/bhagavadgitadu22/Culling_tool_CIP).

Results

1. Scores of the ranking parameters

A scoring system to facilitate effective ranking based on predefined parameters was developed and applied to groups. Two scores were calculated per each group: one emphasizing metadata-based parameters, denoted as SM (score_metadata), and the other concentrating on analytical parameters, referred to as SA (score_analytical). The overall score, labeled as SO (score_overall), is the sum of SM and SA scores.

In practice, each parameter received a score based on the perceived importance of each value, with a minimum value of 0 and a maximum value of 10. This score was binary, i.e. a strain either obtained all the points for a given parameter or none. The sole exception was the parameter "Easiness of replenishment", which allowed for a range of score values to accommodate varying degrees of complexity in bacterial strain culturing. For this particular parameter, eight points were assigned in the absence of any laboratory processing complexity. Conversely, two points were deducted for each requirement introducing complexity, such as culture time exceeding one week, incubation temperature beyond the range of 30°C – 40°C, non-aerobic incubation atmosphere, or the use of a culture medium other than the standard Trypticase Soy Agar medium. The detailed scoring system is shown in Table 1.

In the scope of prioritization for WGS, the global SM score spanned from 0 to 93 points, whereas the global SA score ranged from 0 to 16 points. To emphasize the significance of metadata-based parameters, a relative weight of 6 to 1 was maintained in favor of the SM when calculating the overall score. This decision aimed to prioritize potentially interesting strains, even in cases where these strains had not undergone any analysis, e.g. some of the extremophiles were prioritized even if they had no

associated laboratory analytical data. Following the calculation, all scores were normalized to a scale of 0 to 100 to facilitate the interpretation of results and ensure a consistent base for comparison.

2. Assignment of strains to the groups

The ranking tool was applied to all the unsequenced CIP strains (n=21,314), except for the *Escherichia coli* strains (n=909), as these were being sequenced in the context of an ancillary project and hence were intentionally left out of the scope of the tool. The 20,405 strains were assigned to the 14 groups presented above. Several strains were assigned to more than one group.

Most groups could be built with simple SQL queries extracting the strains matching a list of taxa (archaea, WHO priority pathogens, foodborne zoonotic pathogens, WDCM strains, natural compound producers listed in the NPAtlas database)[20] or the strains with a metadata parameter of a specific value (boolean fields with a true value for the type strains group, or text field with a specific wording for the origin-based subgroups). To identify the type strain species absent from other bacterial biobanks or never sequenced before, we used the gcType database (https://gctype.wdcm.org/, accessed January 2023)[21] managed by the GCM (Global Catalog of Microorganisms, https://gcm.wdcm.org/). The website listed 731 type strain species never sequenced among the 19,440 species with valid published names found at LPSN (List of Prokaryotic names with Standing in Nomenclature; https://lpsn.dsmz.de/, accessed January 2023).

The definition of the two extremophilic subgroups was done as follows: (i) the subgroup "Certain_extremophiles" was based on the definition of extremophiles found in the literature[22–24] and on phenotypic characteristics observed at the CIP, such as acidophilic and alkalinophilic strains growing in pH lower than 5 or higher than 9, halophilic strains requiring a medium with more than 8.8% salt, psychrophilic and thermophilic strains requiring incubation temperatures below 20°C or above 45°C, fitting the criteria reported in Merino et al.[25]; (ii) the subgroup "Uncertain_extremophiles" was set with strains not included in the previous subgroup, but annotated with metadata terms that are frequently associated with extremophilic strains (Supplementary Table S1).

The group definition would have to be rethought to suit ranking purposes of other collections. A specialized collection for instance would have to create more specific groupings to apply this strategy.

3. Compilation of the scores per group and final ranking

Additionally, each group of strains was further categorized according to the acquisition mode, as applicable: (i) Bacterial strains acquired from deposits: this consists of 15,184 strains. These strains have been added to the collection over a long period of time, through the routine operations of the biobank; (ii) Strains inherited from the former Entomopathogenic Bacteria laboratory (IEBC) of the

Pasteur Institute, including 3,100 strains from the taxon *Bacillus* and associated genera that were acquired when the IEBC closed in 2009. At the end, this resulted in 20 groups of strains (Table 2). The next step involved calculating the three scores (SM, SA and SO) for each of the groups. Applying SQL scripts on the PostGreSQL database employed for CIP data management facilitated these score calculations.

The obtained scores per each group are shown in Table 2 (detailed scores for each culling parameter can be found in the Supplementary Table S2). The computed overall scores ranged from 31.3 to 48.0, with a median score of 43.1. Among these groups, the top 10 with the highest scores (above 43.8) were tagged for sequencing, encompassing 9,655 strains, approximately half of the strains initially considered.

To refine the priority list to about 3,000 strains, additional criteria were applied group by group. First, all subgroups with less than 200 strains were prioritized for WGS. These included strains relevant for quality control purposes (n=64), unsequenced type strains (n=76), and uncertain extremophiles (n=197). Strains producing pigments and/or indole were also selected in their entirety due to their moderate number (n=612 and n=471, respectively) and their potential biotechnological value.

Despite the granularity of our ranking methodology, the size of some of the high-scoring groups was still too large for cost efficient WGS purposes, for example those comprising strains isolated from eukaryotes (all strains = 7,536), or soil (only IEBC strains = 1,142).

To reduce the list to make it more cost efficient, we performed an intersection analysis between groups. The UpSet graph[26,27] depicts the size of each group as well as the size of the intersections between the different groups (Figure 3). Then, further categorization was applied based on the intersections between subgroups of interest (e.g. WHO pathogens and potential natural compound producers) to identify higher priority subsets of strains, based on higher cumulative "value".

A final list of 2,871 strains were selected for WGS in the short-term. These strains were prioritized based on their potential applications and relevance to the strategy of the CRBIP and the Pasteur Institute. The selected strains depict a wide diversity in terms of taxonomy, geography, and ecology (Figure 4A-C). Moreover, the genome sequences of only 1,2% of the selected strains were already present in the NCBI database, indicating potential scientific value in the WGS of the other 98,8% prioritized strains. Following WGS, an in-depth bioinformatics analysis is foreseen for each individual species, represented by several individual strains.

Conclusion

We propose a quantitative scoring-based method to efficiently assess and categorize samples in stock in microbial biobanks (Figure 5). This approach can serve different objectives, including identification of high priority strains for sequencing, phenotyping, or identification of strains for culling. In this study, we deployed this method in the scope of strain identification for sequencing. This methodology is based

on the definition of criteria, such as genetic diversity, functional traits, ecological relevance, biomedical potential, or commercial value. The strains with the higher scores correspond to greater scientific and/or commercial value, whereas the strains with the lowest scores could be considered for culling or transferring to other structures.

The proposed methodology has broader applicability and can be extended to other microorganisms, biobanks and for any ranking needs. By adapting and implementing this approach, microbial biobanks stand to enhance management of their collections, with "culling" being a strategic measure towards sustainability. Here, culling is defined as the "process of selectively gathering", allowing biobanks to optimize their resources and contribute to the long-term viability and relevance of their microbial collections.

Author Notes

Two supplementary tables are available with the online version of this article. SQL, Python and R scripts mentioned in the article are available on GitHub at github.com/bhagavadgitadu22/Culling_tool_CIP.

Authorship contribution

MB: conceptualization, formal analysis, investigation, methodology, data curation, visualization, writing – review & editing. AC: conceptualization, formal analysis, investigation, methodology, visualization, writing – original draft, review & editing. MLF: conceptualization, visualization, validation, writing – review & editing. OC, DC: methodology, resources, validation. FB: conceptualization, methodology, supervision, validation, project administration, writing – review & editing.

Conflict of interest

The authors declare no conflict of interest.

Funding

The work was financed by H2020-INFRADEV-03-2019 Project IS_MIRRI21 (Grant Agreement No. 871129).

References

1. Rani A, Saini KC, Bast F, et al. Microorganisms: A Potential Source of Bioactive Molecules for Antioxidant Applications. Molecules 2021, 26(4) ; doi: 10.3390/molecules26041142.
2. Ciani M, Lippolis A, Fava F, et al. Microbes: Food for the Future. Foods 2021, 10(5):971; doi: 10.3390/foods10050971.
3. Priyadarshanee M, Das S. Biosorption and removal of toxic heavy metals by metal tolerating bacteria for bioremediation of metal contamination: A comprehensive review. J Environ Chem Eng 2021, 9(1):104686; doi: 10.1016/J.JECE.2020.104686.
4. Stackebrandt E. Diversification and focusing: Strategies of microbial culture collections. Trends Microbiol 2010, 18(7):283–7; doi: 10.1016/j.tim.2010.05.001.
5. Overmann J, Abt B, Sikorski J. Present and Future of Culturing Bacteria. Annu Rev Microbiol 2017, 71:711-730; doi: 10.1146/annurev-micro-090816-093449.
6. Rackoff LA, Bok K, Green KY, Kapikian AZ. Epidemiology and Evolution of Rotaviruses and Noroviruses from an Archival WHO Global Study in Children (1976–79) with Implications for Vaccine Design. PLoS One 2013, 8(3); doi: 10.1371/journal.pone.0059394.
7. Kapikian AZ, Wyatt RG, Dolin R, et al. Visualization by Immune Electron Microscopy of a 27-nm Particle Associated with Acute Infectious Nonbacterial Gastroenteritis. J Virol 1972, 10(5):1075–81; doi: 10.1128/JVI.10.5.1075-1081.1972.
8. Overmann J. Significance and future role of microbial resource centers. Syst Appl Microbiol 2015 38(4):258–65. Doi: 10.1016/j.syapm.2015.02.008.
9. McCluskey K. A Review of Living Collections with Special Emphasis on Sustainability and Its Impact on Research Across Multiple Disciplines. Biopreserv Biobank 2017, 15(1):20-30; doi: 10.1089/bio.2016.0066.
10. Stackebrandt E, Smith D, Casaregola S, et al. Deposit of microbial strains in public service collections as part of the publication process to underpin good practice in science. Springerplus 2014, 3(1):208; doi: 10.1186/2193-1801-3-208.
11. Cadigan RJ, Juengst E, Davis A, Henderson G. Underutilization of specimens in biobanks: an ethical as well as a practical concern? Genet Med 2014, 16(10):738-40; doi: 10.1038/gim.2014.38.
12. Bledsoe MJ, Sexton KC. Ensuring Effective Utilization of Biospecimens: Design, Marketing, and Other Important Approaches. Biopreserv Biobank 2019, 17(3):248–57; doi: doi: 10.1089/bio.2019.0007.
13. Baláž V, Jeck T, Balog M. Economics of Biobanking: Business or Public Good? Literature Review, Structural and Thematic Analysis. Social Sciences 2022, 11(7):288; doi: 10.3390/SOCSCI11070288.
14. Rush A, Matzke L, Cooper S, et al. Research Perspective on Utilizing and Valuing Tumor Biobanks. Biopreserv Biobank 2019, 17(3):219–29; doi: 10.1089/bio.2018.0099.

15. Snapes E, De Wilde A, Carpenter J, et al. Best Practices: Recommendations for Repositories. 5th Edition. ISBER; Vancouver, BC, 2023.
16. Regulation (EU) No 511/2014 of the European Parliament and of the Council of 16 April 2014. Official Journal of the European Union. 2014. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014R0511 [Last accessed: 12/15/2023].
17. Tacconelli E, Carrara E, Savoldi A, et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. Lancet Infect Dis 2018, 18(3):318–27; doi: 10.1016/S1473-3099(17)30753-3.
18. Koutsoumanis K, Allende A, Alvarez-Ord A, et al. SCIENTIFIC OPINION Role played by the environment in the emergence and spread of antimicrobial resistance (AMR) through the food chain. EFSA J 2021, 19(6):e06651; doi: 10.2903/j.efsa.2021.6651.
19. Ma Q, Zhang X, Qu Y. Biodegradation and Biotransformation of Indole: Advances and Perspectives. Front Microbiol 2018, 9:2625 ; doi: 10.3389/fmicb.2018.02625.
20. Van Santen JA, Poynton EF, Iskakova D, et al. The Natural Products Atlas 2.0: A database of microbially-derived natural products. Nucleic Acids Res 2022, 50(D1):D1317–23; doi: 10.1093/nar/gkab941.
21. Shi W, Sun Q, Fan G, et al. gcType: a high-quality type strain genome database for microbial phylogenetic and functional research. Nucleic Acids Res 2021, 49(D1):D694–705; doi: 10.1093/nar/gkaa957.
22. Grant WD. Life at low water activity. Philos Trans R Soc Lond B Biol Sci 2004, 359(1448):1249-66; discussion 1266-7; doi: 10.1098/rstb.2004.1502.
23. Fang J, Zhang L, Bazylinski DA. Deep-sea piezosphere and piezophiles: geomicrobiology and biogeochemistry. Trends Microbiol 2010, 18(9):413–22; doi: 10.1016/j.tim.2010.06.006.
24. Nogi Y. Taxonomy of Psychrophiles. In: Extremophiles Handbook (Horikoshi K ed.). Springer, Tokyo; 2011.
25. Merino N, Aronson HS, Bojanova DP, et al. Living at the extremes: Extremophiles and the limits of life in a planetary context. Front Microbiol 2019, 10:780; doi: 10.3389/fmicb.2019.00780.
26. Lex A, Gehlenborg N, Strobelt H, et al. UpSet: Visualization of intersecting sets. IEEE Trans Vis Comput Graph 2014, 20(12):1983–92; doi: 10.1109/TVCG.2014.2346248.
27. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 2017, 33(18):2938–40; doi: doi: 10.1093/bioinformatics/btx364.

Figures

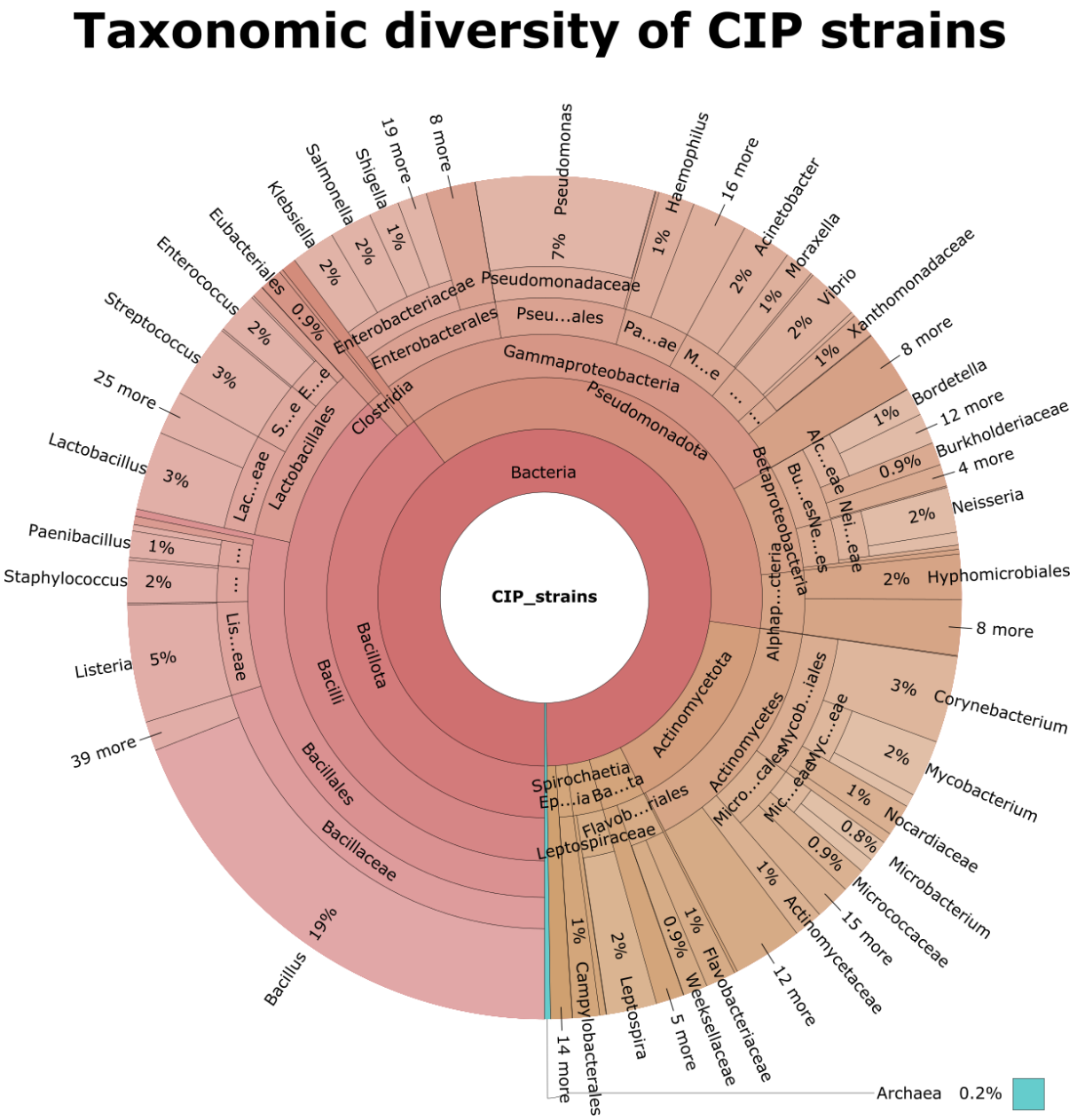

Figure 1. Taxonomic diversity of CIP strains. The CIP hosts about 25,000 strains, of which 99,8% are bacterial strains representing more than 1,000 genera and 5,000 species.

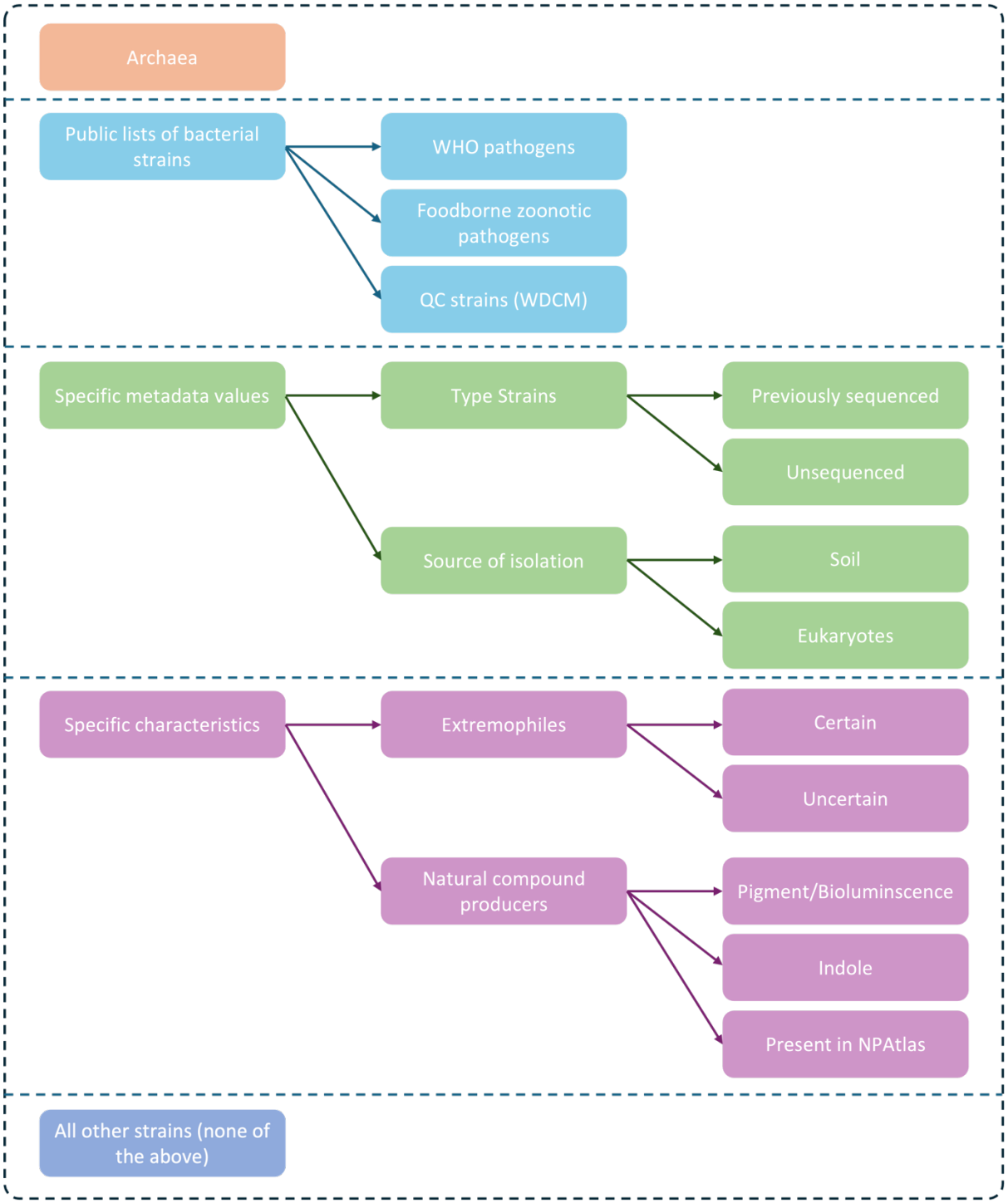

Figure 2. Definition of groups and subgroups of strategic interest.

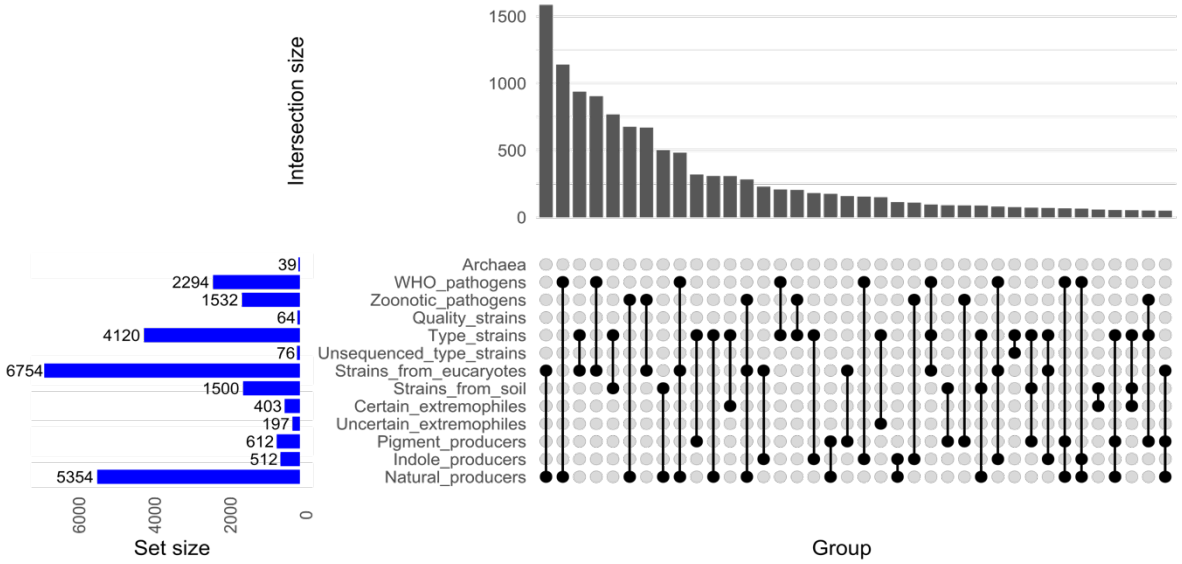

Figure 3. UpSet plot showing the intersections between the groups in a matrix. The rows correspond to the groups, with the number of strains belonging to each group indicated in the left bars, and the columns correspond to the intersections between these groups.

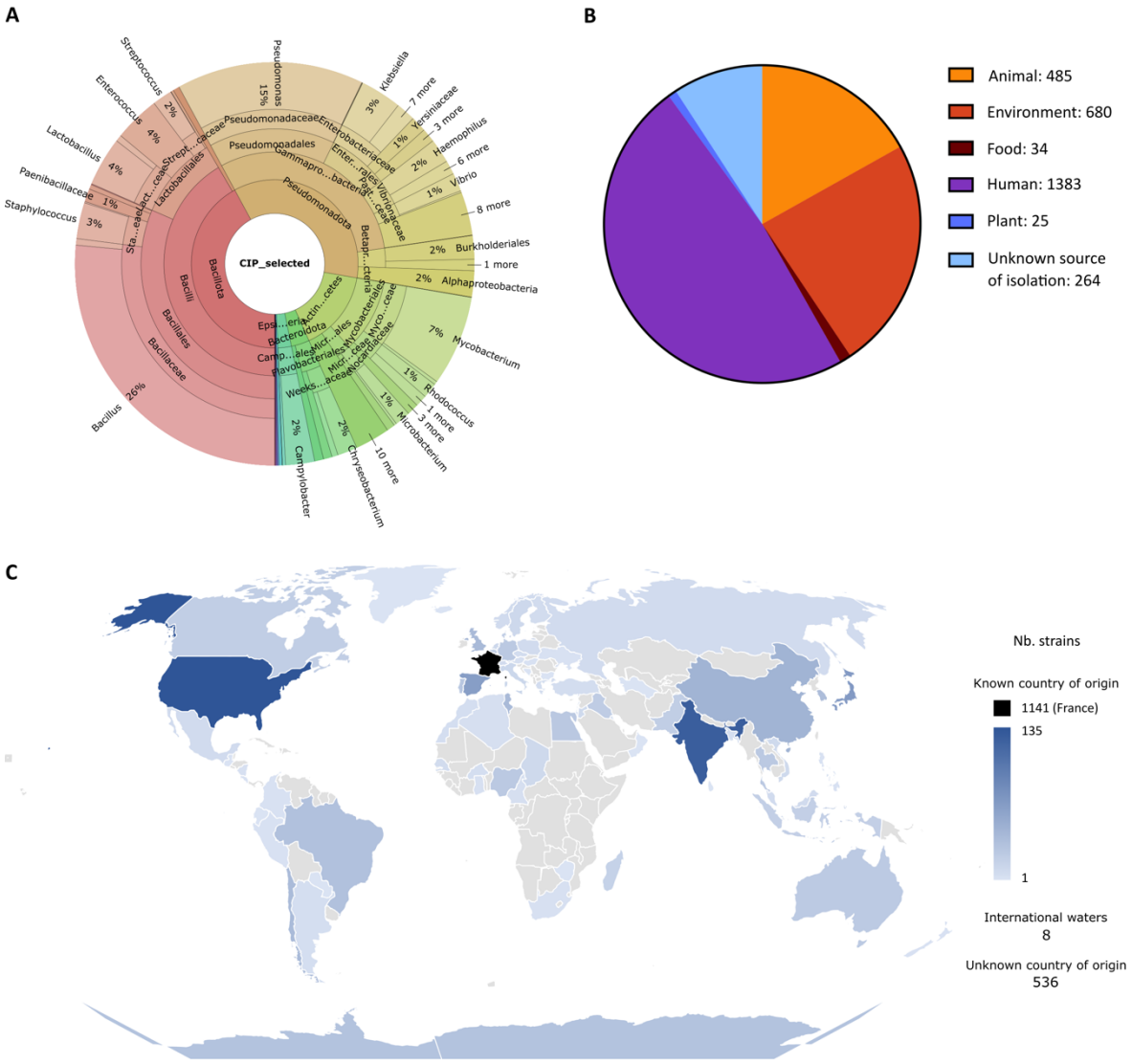

Figure 4. A. Taxonomic diversity of selected CIP strains. B. Source of isolation of selected strains. C. Geographic origin of selected strains.

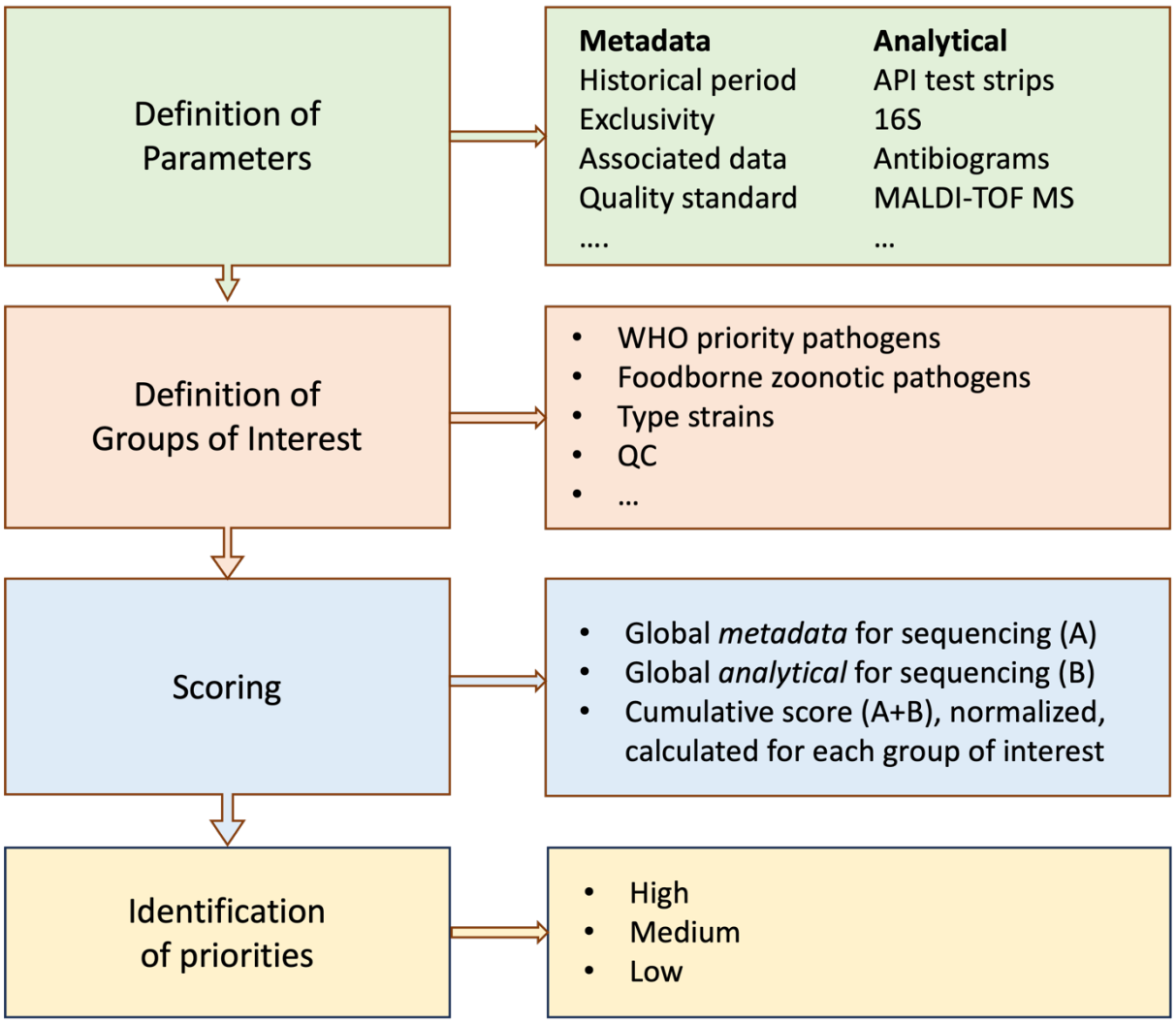

Figure 5. Synthesis of the ranking tool methodology.

Type of information	Parameter	Values	Score	Method of calculation
Metadata score (/93 for the selection of WGS candidates)	Historical period (expressed in percentage of strains)	>35% before 1950	6	For an entire group, score based on the total or the relative number of strains
		>35% between 1950 and 2014	0	
		>35% after 2014	9	
	Number of strains (per group)	0-10	5	
		10-100	2	
		100-1000	0	
	Distribution rate (5 last years)	>50%	3	
		10-50%	2	
		1-10%	1	
		<1%	0	
	Status in BIMS and catalog	integrated in the BIMS & available in the catalog	10	By strain, all results summed and divided by the number of strains to get the average result
		non-integrated in the BIMS & not available in the catalog	5	
	Exclusivity (present/absent in other collections)	present	0	
		absent	8	
	Number of distribution samples in stock (aliquots)	0	0	
		1-10	2	
		10-100	6	
		100-500	8	
	Age of the most recent distribution batch (years)	less than 5 years	4	
		more than 5 years	2	
		undefined	0	
	Storage temperature	-20°C	0	
		-80°C	2	
		LN	3	
		RT	1	
	WDCM standard	WDCM yes	3	
		WDCM no	0	
	Associated data	year	8	
		geographic location	8	
		source	8	
	Collection replenishment	easiness of replenishment	8-2*number-of-complications	

	Viability data	less than 10 years	2	
		more than 10 years	0	
		undefined	0	
Analytical score (/16 for the selection of WGS candidates)	API test strips	done	2	By strain, all results summed and divided by the number of strains to get the average result
	16S rRNA or housekeeping genes	done	5	
	Antibiograms	done	5	
	MALDI-TOF MS	done	4	

Table 1. List of parameters used to evaluate the sequencing pertinence of each group, and scoring system used to classify the groups. The first column establishes the separation between metadata criteria and analytical criteria.

Group	IEBC subgroup	Number of strains unsequenced by CIP	Metadata score (SM) of strains (/100)	Analytic score (SA) of strains (/100)	Global score (SO) of strains (/100)	% of WGS unavailable on NCBI
Archaea	no	39	42,4	20,4	39,2	30,8
WHO_pathogens	no	2294	44,0	29,7	41,9	90,0
Zoonotic_pathogens	no	1532	42,9	17,6	39,2	91,4
Quality_strains	no	64	41,1	85,7	47,7	48,4
Type_strains	no	4119	42,9	44,6	43,1	45,8
Unsequenced_type_strains	no	76	44,8	39,5	44,0	98,7
Strains_from_eucaryotes	no	6297	50,2	29,4	47,2	89,3
Strains_from_eucaryotes	yes	457	51,8	9,4	45,6	100,0
Strains_from_soil	no	944	42,8	39,8	42,4	55,4
Strains_from_soil	yes	555	53,0	7,6	46,3	100,0
Certain_extremophiles	no	403	41,4	32,1	40,1	58,6
Uncertain_extremophiles	no	176	44,2	41,3	43,8	52,8
Uncertain_extremophiles	yes	21	53,1	4,2	45,9	100,0
Pigment_producers	no	612	42,8	54,4	44,5	67,3
Indole_producers	no	471	47,3	51,9	48,0	75,8
Indole_producers	yes	41	47,4	12,5	42,3	100,0
Natural_producers	no	3351	43,4	27,7	41,1	90,0
Natural_producers	yes	2003	50,3	10,2	44,4	100,0
Other_strains	no	2653	38,3	19,1	35,5	97,3
Other_strains	yes	682	41,7	5,6	36,4	100,0

Table 2. Score obtained per group with a separation between the strains from the former IEBC laboratory and the other strains. The third column details the number of strains unsequenced by CIP at the time of writing, the following columns details the reasons why those strains should be sequenced: metadata, analytic and global score of the strains as well as percentage of strains unsequenced according to NCBI data. The colors represent a hit map that spans from green to red, with green being the highest value and red the lowest value for each column.

Lexic	Likeliness
acid	yes
acidic	yes
alkaline	yes
altitude	no
anaerobic	no
anoxic	yes
Antarctic	yes
Antarctica	yes
arctic	yes
arid	yes
Atacama	yes
biofilm	no
biofilms	no
brine	no
cave	yes
cold	no
deep	yes
dehydrated	yes
depth	no
desert	yes
desiccation	yes
dry	no
dry-heated	yes
frozen	yes
fumarole	yes
geothermal	yes
glacier	yes
high	no
highly	no
hot	no
hot spring	yes
hydrothermal	yes
hypermagnesian	yes
hypersaline	yes
ice	yes
irradiated	yes
light	no
manure	no
Mariana trench	no
mine	yes
nuclear	yes
ophiolite	yes
oven	yes

permafrost	yes
pH	no
polar	yes
radiation	yes
saline	no
salinity	no
salt	no
salted	no
salty	no
serpentinite	yes
soda	yes
International Space Station	yes
subantarctic	yes
temperature	no
trench	no
vent	no
volcano	yes
warm	no
Yellowstone	no

Table S1. List of extremophilic terms used to build the subgroup "Uncertain extremophiles".

Group	IEBC subgroup	Count	Historical value (/16)	Number of strains (/5)	Distribution rate (/3)	Integration status (/10)	Exclusivity (/8)	Number of distribution samples in stock (/8)	Age of the latest distribution batch (/4)	Type of storage (/3)	WDCM (/3)	Year known (/8)	Country known (/8)	Origin known (/8)	Easiness of replenishment (/8)	Viability (/2)	Metadata score (/93)	API (/2)	WGS (/10)	16S rRNA or housekeeping genes (/5)	Antibiograms (/5)	MALDI-TOF (/4)	Analytic score (/16)	Global score (/109)
Archaea	no	39	0	2	0	8,59	0,21	2,46	1,85	0,72	0	2,87	7,38	7,38	5,9	0,1	39,46	1,23	0	0,26	1,67	0,1	3,26	42,72
WHO_pathogens	no	2294	0	0	2	8,17	6,35	2,23	1,62	0,54	0,02	4,3	4,51	3,98	6,86	0,35	40,91	0,94	0	1,71	1,75	0,33	4,75	45,66
Zoonotic_pathogens	no	1532	0	0	1	6,6	6,19	1,24	1,22	0,28	0,02	5,14	4,95	5,44	7,49	0,34	39,89	0,6	0	1,03	0,9	0,27	2,82	42,71
Quality_strains	no	64	0	2	3	10	0	4	2,09	0,92	3	1,75	1,63	5,5	6,34	1	38,23	1,75	0	4,61	2,66	1,69	13,71	51,94
Type_strains	no	4119	0	0	2	9,47	0,38	2,55	1,92	0,89	0,03	3,07	6,14	7,49	5,77	0,2	39,88	1,57	0	3,16	2,06	0,32	7,14	47,02
Unsequenced_type_strains	no	76	0	2	1	9,08	0,32	2,5	1,92	0,82	0	3,16	7,26	7,79	5,66	0,16	41,67	1,45	0	2,7	1,91	0,26	6,32	47,99
Strains_from_eucaryotes	no	6297	0	0	2	7,96	6,22	2,15	1,36	0,58	0,01	6,25	6,61	7,16	6,23	0,16	46,68	0,95	0	1,82	1,59	0,34	4,71	51,39
Strains_from_eucaryotes	yes	457	0	0	1	7,93	7,51	0,36	0,43	0,12	0	7,19	7,54	7,93	8	0,2	48,21	1,36	0	0	0,14	0	1,5	49,71
Strains_from_soil	no	944	0	0	1	9,34	0,74	2,38	1,95	0,93	0	2,42	6,52	8	6,37	0,16	39,81	1,55	0	2,64	1,92	0,26	6,37	46,18
Strains_from_soil	yes	555	0	0	2	7,32	7,87	0,35	0,69	0,15	0	6,49	8	8	8	0,39	49,26	1,11	0	0	0,11	0	1,22	50,48
Certain_extremophiles	no	403	0	0	1	9	1,31	2,77	1,88	0,76	0,01	3,2	6,25	7,23	5,02	0,11	38,53	1,33	0	2	1,66	0,13	5,13	43,66
Uncertain_extremophiles	no	176	0	0	1	9,15	0,91	2,31	1,92	0,84	0,02	3,64	7,55	7,91	5,69	0,18	41,1	1,72	0	2,76	1,68	0,43	6,61	47,71
Uncertain_extremophiles	yes	21	0	2	0	6,9	8	0,29	0,1	0	0	8	8	8	8	0,1	49,39	0,67	0	0	0	0	0,67	50,06
Pigment_producers	no	612	0	0	2	9,61	2,22	2,81	1,96	0,84	0,02	3,35	4,88	5,95	6,07	0,1	39,79	1,87	0	2,34	4,26	0,21	8,7	48,49
Indole_producers	no	471	0	0	2	9,69	3,82	3,15	1,97	0,9	0,04	4,25	5,54	6,06	6,4	0,2	43,98	1,95	0	3,22	2,61	0,49	8,31	52,29
Indole_producers	yes	41	0	2	0	9,88	7,22	0	0	0	0	8	8	0,98	8	0	44,08	2	0	0	0	0	2	46,08
Natural_producers	no	3351	0	0	2	7,91	6,21	1,97	1,52	0,47	0,03	4,27	4,46	4,65	6,66	0,28	40,4	0,88	0	1,59	1,57	0,36	4,43	44,83
Natural_producers	yes	2003	0	0	2	7,46	7,65	0,35	0,62	0,15	0	7,95	7,56	4,66	8	0,33	46,73	1,49	0	0	0,14	0	1,63	48,36
Other_strains	no	2653	0	0	1	7,2	6,5	1,64	1,16	0,49	0	4,33	3,26	3,37	6,56	0,09	35,6	0,77	0	1,23	0,92	0,13	3,05	38,65
Other_strains	yes	682	0	0	0	7,68	7,5	0	0	0	0	5,87	7,62	2,08	8	0	38,75	0,89	0	0	0	0	0,89	39,64

Table S2. Detailed scores per criteria obtained for each group with a separation between the strains from the former IEBC laboratory and the other strains. The scores are absolute contrary to the table of scores from the main text.